frontiers in Psychology

HYPOTHESIS AND THEORY
published: 27 December 2016
doi: 10.3389/fpsyg.2016.02019

A Bayesian Framework for False Belief Reasoning in Children: A Rational Integration of Theory-Theory and Simulation Theory

Nobuhiko Asakura and Toshio Inui*

Department of Psychology, Otemon Gakuin University, Osaka, Japan

Two apparently contrasting theories have been proposed to account for the development of children's theory of mind (ToM): theory-theory and simulation theory. We present a Bayesian framework that rationally integrates both theories for false belief reasoning. This framework exploits two internal models for predicting the belief states of others: one of self and one of others. These internal models are responsible for simulation-based and theory-based reasoning, respectively. The framework further takes into account empirical studies of a developmental ToM scale (e.g., Wellman and Liu, 2004): developmental progressions of various mental state understandings leading up to false belief understanding. By representing the internal models and their interactions as a causal Bayesian network, we formalize the model of children's false belief reasoning as probabilistic computations on the Bayesian network. This model probabilistically weighs and combines the two internal models and predicts children's false belief ability as a multiplicative effect of their early-developed abilities to understand the mental concepts of diverse beliefs and knowledge access. Specifically, the model predicts that children's proportion of correct responses on a false belief task can be closely approximated as the product of their proportions correct on the diverse belief and knowledge access tasks. To validate this prediction, we illustrate that our model provides good fits to a variety of ToM scale data for preschool children. We discuss the implications and extensions of our model for a deeper understanding of developmental progressions of children's ToM abilities.

Keywords: false belief, theory-theory, simulation theory, Bayesian network, internal model

Edited by:
Erika Nurmsoo,
University of Kent, UK

Reviewed by:
Caspar Addyman,
Goldsmiths, University of London, UK
Yoshifumi Ikeda,
Joetsu University of Education, Japan

****Correspondence:***
Nobuhiko Asakura
n-asakura@otemon.ac.jp

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *21 July 2016*
Accepted: *12 December 2016*
Published: *27 December 2016*

Citation:
Asakura N and Inui T (2016) A Bayesian Framework for False Belief Reasoning in Children: A Rational Integration of Theory-Theory and Simulation Theory. Front. Psychol. 7:2019. doi: 10.3389/fpsyg.2016.02019

1. INTRODUCTION

Inferring and understanding other people's mental states such as desires, beliefs, and intentions is crucial for our successful social interactions. This ability has been referred to as having a "theory of mind" (ToM; Premack and Woodruff, 1978). For decades, ToM development in childhood has been the subject of intensive research; much of the research has focused on children's false belief understanding. Two false belief tasks are widely used for assessing children's ToM: unexpected-contents (Perner et al., 1987) and change-of-location (Wimmer and Perner, 1983) tasks. In

the unexpected-contents task, children are shown a familiar container that holds something unexpected inside when it is opened. They are then asked about what an agent will think is inside the container when she has never seen it opened. In the change-of-location task, children are shown a situation in which an agent places an object at one location, leaves the scene, and then her antagonist moves it to another location while she is gone. They are then asked about where the agent will look for the object after she returns. Correct answers for these tasks require children to appreciate that an agent can have a false belief that contradicts the reality with which they are faced. Hence, success on false belief tasks is taken as indicating that ToM has become mature enough to function as an inference engine for reasoning about other people's beliefs, as distinct from one's own.

Many studies have revealed that children come to understand other people's false beliefs at around 4 or 5 years of age; in addition, such a developmental transition appears to occur gradually (e.g., Wellman et al., 2001, for a review). Two main theories have been proposed for explaining the process of ToM development: theory-theory (Gopnik and Wellman, 1992, 1994) and simulation theory (Gordon, 1986; Gallese and Goldman, 1998). Theory-theory assumes that ToM ability rests on a set of rules, or literally theories, about how the minds of others work. It thus claims that children learn and become able to use such theories to predict and explain others' mental states and their behavior. In contrast, simulation theory argues that ToM ability does not require theorizing the minds of others. Instead, it claims that children come to use their own minds as a simulation model to mimic and understand the minds of others.

These theories have long been regarded as contrasting conceptualizations of ToM. In recent years, however, a number of researchers have advocated hybrid theories that incorporate the essences of both theory and simulation (Nichols and Stich, 2003; Saxe, 2005; Goldman, 2006; Mitchell et al., 2009). Notably, Mitchell et al. (2009) proposed that children first acquire a competence of simulation and then develop a theory-based reasoning skill, and they will adopt both of these reasoning strategies depending on the demands of a particular task at hand. Recent neuroimaging findings further support such hybrid approaches, demonstrating mixed evidence for the neural mechanisms of ToM responsible for either theory-based or simulation-based reasoning (Apperly, 2008; Mahy et al., 2014).

In spite of a large body of empirical findings and recent theoretical advances in ToM research, relatively few studies have proposed computational models of ToM understanding, particularly false belief understanding (O'Laughlin and Thagard, 2000; Goodman et al., 2006; Berthiaume et al., 2013). Moreover, none deal with mixed reasoning strategies based on theory-theory and simulation theory. Therefore, it still remains unclear whether and how children can, in principle, combine both strategies into a coherent theory of mind. In this study, we present a computational model that integrates theory-based and simulation-based strategies for false belief reasoning. Our model builds on a Bayesian framework and thus provides a rational account of children's ToM. It also makes testable predictions about children's performance on false belief tasks, allowing a quantitative comparison with existing behavioral data.

We argue that a developmental ToM scale (Wellman and Liu, 2004) is of particular relevance for any computational model of false beliefs. The ToM scale consists of tasks to assess children's understanding of multiple mental state concepts. It reflects extant findings of children's ToM such that they develop an understanding of diverse desires (people can have different desires for the same thing) before developing that of diverse beliefs (people can have different opinions and beliefs about the same situation); they develop understandings of diverse beliefs and knowledge access (others can have different perspectives that prevent them from having access to the true real-world information) before developing that of false beliefs. This kind of developmental sequence has been confirmed for preschool children with diverse cultural backgrounds (Wellman and Liu, 2004; Peterson et al., 2005; Wellman et al., 2006; Toyama, 2007; Shahaeian et al., 2011; Hiller et al., 2014). From a constructivism point of view, such a sequential progression of ToM suggests that an understanding of false beliefs should emerge under the developed understandings of the mental concepts such as diverse desires, diverse beliefs, and knowledge access. Taking into account this view, we formalize a model of false belief reasoning based on a Bayesian network (Pearl, 2000; Spirtes et al., 2001) that represents causal relationships among the relevant mental concepts of others and one's own. In so doing, we show that this Bayesian network in effect provides a natural way to integrate theory-based and simulation-based strategies. We further demonstrate that our model provides a good fit to the existing ToM scale data.

2. MODEL

2.1. Bayesian Network

A Bayesian network is a graphical model that provides a compact representation of the joint probability distribution for a set of random variables (Pearl, 2000; Spirtes et al., 2001). Its graph structure represents the causal probabilistic relationship among the variables, specifies a particular factorization of the joint probability distribution, and enables efficient computation of probability distributions of the unobserved variables, given the observed ones. Bayesian networks have been used in a wide range of fields and applications, such as computer science, engineering, statistics, medical diagnosis, and bioinformatics. Recently, they have found application in various areas of psychology, such as visual perception (Kersten et al., 2004), cognition (Griffiths et al., 2008; Jacobs and Kruschke, 2011), causal inference (Griffiths and Tenenbaum, 2005; Lu et al., 2008), and cognitive development (Gopnik et al., 2004; Gopnik and Tenenbaum, 2007; Gopnik and Wellman, 2012). Notably, Gopnik and Wellman (2012) have argued that Bayesian networks can be used for formalizing a theory-theory of cognitive development. In the following, we use this Bayesian network formalism to elaborate a model of false belief reasoning in children.

From a theory-theory perspective, Goodman et al. (2006) proposed Bayesian network models of false belief reasoning in the change-of-location task. Our Bayesian formulation closely follows their work; the differences are that we additionally take into account the idea of simulation theory and that we focus on

the unexpected-contents task to model false belief reasoning. The latter is motivated by the fact that this type of task was commonly used across the ToM scale studies listed above, but has not been the subject of formal analysis. The extension of our model to the change-of-location task will be discussed later.

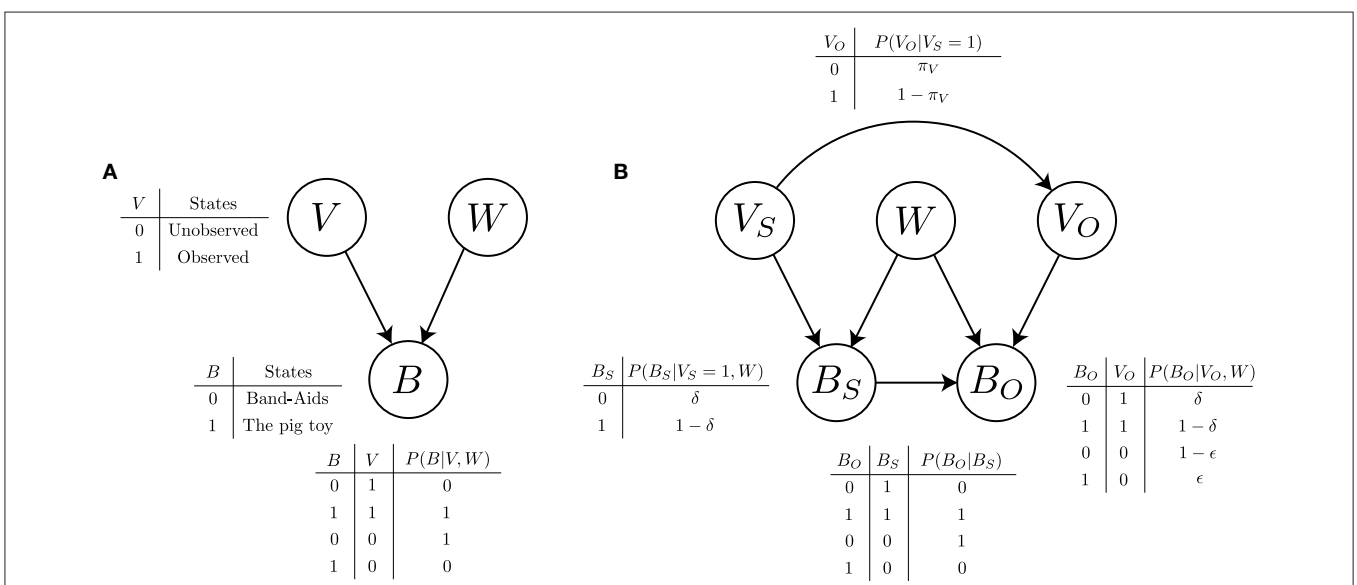
We note that the unexpected-contents task is divided into two stages. According to Wellman and Liu (2004), in the first stage, a child sees a familiar, closed Band-Aid box that holds inside a plastic pig toy. From the appearance of the Band-Aid box, the child first expects Band-Aids inside. Subsequently, the Band-Aid box is opened and the child sees the pig toy inside. Then, the Band-Aid box is closed again. In the second stage, there is a toy figure of a boy named Peter. The child hears that Peter has never seen inside this Band-Aid box. Then, the child answers the question: What does Peter think is in the box? Band-Aids or a pig? Thus for the child, the first stage concerns updating the belief state of the self, whereas the second stage involves reasoning about the belief state of others.

These consecutive stages can be formalized within a Bayesian network framework as follows. Consider the process of belief updating in the first stage. The initial belief about the hidden contents of the Band-Aid box (i.e., Band-Aids) comes from observing the outside of the box. Then, the updated belief (i.e., the pig toy) arises from having visual access to the inside of the box. This in turn implies that without such visual access, the initial belief wound not be updated, but instead remain in its original state.

These causal relationships can be concisely represented with a causal graphical model or a Bayesian network, as depicted in **Figure 1A**. This graph has three nodes of random variables (W, V, and B) and two directed edges between the nodes (i.e., single arrows). W represents the true state of the world, that is, the Band-Aid box that holds the pig toy inside. V represents the binary states of visual access to the inside of the box: the contents are observed or not. B represents the binary states of the belief about the contents: the pig toy (true belief) or Band-Aids (false belief). The directed edges represent causal connections between these variables: each arrow points from a cause to an effect. Accordingly, this graph specifies a causal relation such that V and W are the cause of B. In addition, it implies a probabilistic interpretation of the causal relation in terms of the conditional probability of B given V and W: $P(B|V, W)$. To be more specific, this graph defines the joint probability distribution over all variables $P(B, V, W)$, and the causal structure implies a particular factorization of the joint probability distribution as $P(B, V, W) = P(B|V, W)P(V)P(W)$. This Bayesian network thus denotes how belief formation proceeds in children's minds for the unexpected-contents task.

Next, consider belief reasoning in the second stage. Given the updated belief in the first stage and available information about Peter's visual access, the child makes an inference about Peter's belief about the hidden contents of the Band-Aid box. We propose that the child makes a probabilistic inference about Peter's belief using a "theory" that represents the process of belief updating in the first stage. Specifically, we propose that at the computational level (Marr, 1982), the child's belief reasoning can be formalized as probabilistic computations on a Bayesian network that represents the causal process of belief formation (**Figure 1B**).

To construct this Bayesian network, we make two assumptions. First, we assume that the child has two theories of belief formation, one applied to her and the other to Peter, each of which can be represented as the Bayesian network in **Figure 1A**. In other words, we postulate two internal models



V	States
0	Unobserved
1	Observed

B	States
0	Band-Aids
1	The pig toy

B	V	$P(B\|V,W)$
0	1	0
1	1	1
0	0	1
1	0	0

V_O	$P(V_O\|V_S=1)$
0	π_V
1	$1-\pi_V$

B_S	$P(B_S\|V_S=1,W)$
0	δ
1	$1-\delta$

B_O	B_S	$P(B_O\|B_S)$
0	1	0
1	1	1
0	0	1
1	0	0

B_O	V_O	$P(B_O\|V_O,W)$
0	1	δ
1	1	$1-\delta$
0	0	$1-\epsilon$
1	0	ϵ

FIGURE 1 | (A) Bayesian network describing a causal relation among the states of the world (W), visual access (V), and belief (B). The conditional probability table indicates that the belief updating is deterministic. **(B)** Our Bayesian network model for false belief reasoning. The S subscript is the abbreviation for "self;" the O subscript for "others."

(Wolpert et al., 2003) of self and others for belief prediction. In effect, the internal model of self acts as a simulator of one's own mind. This can be viewed as a form of rule-based simulation (cf. Mitchell et al., 2009). Second, we assume that during children's early development, their internal model of self is inseparable from and affects that of others. Specifically, we hypothesize that when reasoning about Peter's belief, the child's own states of belief and visual access have an effect on those in the internal model of Peter. This hypothesis embodies a true-belief default (Leslie et al., 2004): others tend to have the same true belief as one's own, as people usually have true beliefs about everyday matters. It also follows from the "like-me" hypothesis of infant social cognitive development (Meltzoff, 2007a,b).

These two assumptions lead us to construct the Bayesian network depicted in **Figure 1B**. This graph has two sets of random variables: V_S and B_S, representing the binary states of visual access and belief, respectively, of self; and V_O and B_O of others. Given W, each set of the variables can constitute a subgraph that exactly matches the graph in **Figure 1A**. Then, merging the two subgraphs leads to jointly representing two internal models of belief updating: one for the self and one for others. This meets the first assumption. Further, in **Figure 1B**, the whole graph has two extra edges between the variables of the self and others: one from V_S to V_O, and the other from B_S to B_O. This represents their causal connections and thereby reflects the second assumption. Furthermore, this Bayesian network can be interpreted as representing the joint probability distribution over all relevant variables, implying its factorization as $P(V_O, B_O, V_S, B_S, W) = P(B_O|V_O, B_S, W)P(V_O|V_S)P(B_S|V_S, W)P(V_S)P(W)$.

The Bayesian network in **Figure 1B** thus describes our proposed theory of mind that children might use in the unexpected-contents false belief task. From now on, we refer to this Bayesian network as the ToM network. Its use allows children to perform theory-based reasoning about the belief states of others. But this reasoning does not strictly follow the standard theory-theory in which the mental states of others are detached from those of one's own. Rather, reasoning with this ToM network inevitably entails more or less simulation-based reasoning through the internal model of one's own mind. Therefore, our proposed ToM network can be viewed as a hybrid model combining both theory-based and simulation-based reasoning. In the following section, we formulate false belief reasoning as Bayesian inference with such a hybrid model. This can be done through parameterization of the ToM network to specify a conditional probability for each relevant variable.

2.2. False Belief Reasoning

First, we introduce notations to denote the states of a binary variable. For the states of visual access to the inside of the Band-Aid box, let V_S and V_O take on the value of 1 when the contents are observed and 0 otherwise. For the states of belief about the contents, let B_S and B_O take on the value of 1 for the pig toy (true belief) and 0 for Band-Aids (false belief).

Given the ToM network, reasoning about the belief state of others amounts to estimating the state of the variable B_O from available information about the remaining variables. For this estimation, the Bayesian approach suggests using the predicted probability $P(B_O|V_S = 1, W)$: the conditional probability of B_O given the observed state of V_S and the fixed state of W. Recall that the ToM network represents the joint probability distribution over all variables. Then, from this, the predicted probability can be computed by applying Bayes' rule and marginalization:

$$P(B_O|V_S = 1, W) = \sum_{V_O}\sum_{B_S} P(B_O|V_O, B_S, W)P(V_O|V_S = 1) P(B_S|V_S = 1, W) \quad (1)$$

Here, the summation is taken over all possible values of V_O and B_O. Hence, to formulate this predicted probability, it is necessary to specify three conditional probabilities on the right side of the Equation (1).

First consider $P(B_S|V_S = 1, W)$. This conditional probability concerns only the mental states of self and represents children's belief updating in the first stage of the unexpected-contents task. As the child comes to hold the true belief when she observes the inside of the box (i.e., $V_S = 1$), B_S will take on the value of 1 when her belief is just updated. We assume, however, that the child may fail to maintain the updated belief owing to accidental error, or the limited capacity of her working memory. Assuming that such failure occurs with small probability δ, we set

$$P(B_S = 0|V_S = 1, W) = \delta \quad (2)$$

$$P(B_S = 1|V_S = 1, W) = 1 - \delta \quad (3)$$

Next, consider $P(V_O|V_S = 1)$. This conditional probability is due to our assumption that the mental states of self have an effect on the representations of those of others. As the child hears that Peter has not observed the inside of the box, V_O should be 0 if she correctly identifies the state of Peter's visual access. However, our assumption states that the child may mistakenly attribute her state of mind to Peter. Assuming that this happens with probability $1 - \pi_V$, we set

$$P(V_O = 0|V_S = 1) = \pi_V \quad (4)$$

$$P(V_O = 1|V_S = 1) = 1 - \pi_V \quad (5)$$

Thus, π_V expresses the degree to which children can appreciate the states of others' visual access, or equivalently, the knowledge states of others.

Finally, consider $P(B_O|V_O, B_S, W)$. This conditional probability represents the process of others' belief updating with its dependence on the belief states of self. This dependence is again due to our assumption stated above. We assume that the child may automatically adopt her state of belief to represent Peter's own and this occurs with a probability of $1 - \pi_B$. This prompts us to decompose $P(B_O|V_O, B_S, W)$ as follows:

$$P(B_O|V_O, B_S, W) = \pi_B P(B_O|V_O, W) + (1 - \pi_B)P(B_O|B_S) \quad (6)$$

This decomposition implies that to represent Peter's belief, the child uses $P(B_O|V_O, W)$ with probability π_B, or $P(B_O|B_S)$ with probability $1 - \pi_B$. Thus, π_B expresses the degree to which children can attribute different beliefs to others.

Note that $P(B_O|V_O, W)$ concerns only the mental states of others and represents the identical causal structure with $P(B_S|V_S, W)$. Therefore, as Equations (2) and (3), when $V_O = 1$, we set

$$P(B_O = 0|V_O = 1, W) = \delta \quad (7)$$

$$P(B_O = 1|V_O = 1, W) = 1 - \delta \quad (8)$$

When $V_O = 0$, Peter should hold the false belief ($B_O = 0$) because he observes only the outside of the Band-Aid box. However, we assume that the child takes into account the possibility that Peter expects something other than Band-Aids inside the box. This can happen because the box is a container that can hold anything smaller than its size; in fact, it held the pig toy inside in the unexpected-contents task. Assuming that the child supposes such a misconception could occur with a small probability ϵ, we set

$$P(B_O = 0|V_O = 0, W) = 1 - \epsilon \quad (9)$$

$$P(B_O = 1|V_O = 0, W) = \epsilon \quad (10)$$

For $P(B_O|B_S)$, we assume that it is deterministic since its probabilistic nature has already been captured with the probability $1 - \pi_B$. That is, assuming that the child's state of belief is just copied to Peter's belief, we set

$$P(B_O = 0|B_S = 0) = P(B_O = 1|B_S = 1) = 1 \quad (11)$$

$$P(B_O = 0|B_S = 1) = P(B_O = 1|B_S = 0) = 0 \quad (12)$$

By using the parameterization introduced thus far, we can derive the predicted probability $P(B_O|V_S = 1, W)$, and then our model of false belief reasoning as $P(B_O = 0|V_S = 1, W)$. First, we substitute Equation (6) into Equation (1) to obtain:

$$P(B_O|V_S = 1, W) = \pi_B \sum_{V_O} P(B_O|V_O, W)P(V_O|V_S = 1) + (1 - \pi_B) \sum_{B_S} P(B_O|B_S)P(B_S|V_S = 1, W) \quad (13)$$

This illustrates how theory-based and simulation-based strategies are combined to perform reasoning about the belief state of others. The first term means that the child first obtains an estimate of the state of Peter's visual access using her own state ($P(V_O|V_S = 1)$), then feeds the estimate into the internal model of others ($P(B_O|V_O, W)$) to predict the belief state of Peter. This corresponds to a theory-based strategy. In contrast, the second term means that the child first employs the internal model of self ($P(B_S|V_S = 1, W)$) to simulate her own belief updating and then projects the simulated state of belief onto Peter ($P(B_O|B_S)$). This corresponds to a simulation-based strategy. The probability π_B acts as a gate to select one of these strategies: the child performs theory-based reasoning with probability π_B and simulation-based reasoning with probability $1 - \pi_B$.

Then, by setting $B_O = 0$ and summing V_O and B_S in Equation (13), we finally obtain our model of false belief reasoning:

$$P(B_O = 0|V_S = 1, W) = \pi_B\pi_V(1 - \epsilon) + (1 - \pi_B\pi_V)\delta \quad (14)$$

This is the probability that given knowledge about the true state of the world, the child estimates the belief state of Peter as a false belief. In the following, we denote this probability as π_{FB}.

2.3. Relation to the Theory-of-Mind Scale

Our model of false belief reasoning takes four probabilities as its parameters: δ, ϵ, π_B, and π_V. Of these, the effects of δ and ϵ are likely to be limited since they are assumed to be small, random errors. This is justified by the procedure employed in all the ToM scale studies listed above. In fact, for children's false belief responses to be scored as correct, the children were first required to respond correctly to preliminary and control questions about what is usually in a Band-Aid box (i.e., Band-Aids) and what is actually in the Band-Aid box presented (i.e., the pig toy). Thus, we can safely assume that children rarely, if ever, came up with something other than Band-Aids inside the box (i.e., small ϵ) and failed to maintain the updated belief about the contents of the box (i.e., small δ). In contrast, the remaining two probabilities, π_B, and π_V, play a dominant role in specifying the behavior of our model. Let us remember that π_B and π_V are introduced to quantify children's abilities to differentiate their mental states, beliefs, and visual access, respectively, from those of others. These abilities as well as false belief reasoning are, in fact, ToM skills that are to be assessed with the ToM scale (Wellman and Liu, 2004). Two relevant ToM tasks included in the scale: diverse beliefs and knowledge access. A diverse-beliefs task involves the ability to understand that others can have different beliefs about the same situation. A knowledge access task involves the ability to discern others' visual access to judge whether they are knowledgeable or ignorant. Hence, π_B corresponds to the proportion correct for the diverse-beliefs task, and π_V for the knowledge access task. Obviously, π_{FB} amounts to children's proportion correct for the unexpected-contents false-beliefs task.

Our model thus predicts that, when assessed with the ToM scale in terms of the proportion correct, children's false belief ability can be predicted through their abilities to understand diverse beliefs and knowledge access. Indeed, assuming that δ and ϵ are sufficiently small, we can approximate Equation (14) to obtain a simple relation: $\pi_{FB} \approx \pi_B\pi_V$. It follows that false belief reasoning can be viewed as a multiplicative effect of understanding diverse beliefs and knowledge access. Below we will show that this simple multiplicative relation holds across a wide variety of children's ToM scale data.

3. RESULTS

We illustrated the validity of our model by fitting the full model of four parameters using a Bayesian method to the children's ToM scale data for the three above-mentioned tasks. The data for each task consist of the number of children who successfully completed the task. To fit our model to the data, we take the proportion of children who were correct on each task as their proportion correct for the task. Specifically, we assume an individual child's responses to these tasks as independent Bernoulli trials with success probabilities π_B for

diverse beliefs, π_V for knowledge access, and π_{FB} for unexpected-contents false belief. This allows us to derive the joint likelihood function of the parameters π_B, π_V, δ, and ϵ (note that π_{FB} is a function of them). In addition, similar to Goodman et al. (2006), we assume asymmetric beta priors on δ and ϵ to make their small values more likely than large ones. Then, we combine the likelihood and prior to form the posterior distribution over the parameters, from which we find their maximum a posteriori estimates (see the Appendix for details). Given these estimates, we will obtain a prediction of π_{FB} using Equation (14).

Figure 2 shows the comparison of our model prediction with the data from several ToM scale studies (Wellman and Liu, 2004; Peterson et al., 2005; Wellman et al., 2006; Toyama, 2007; Shahaeian et al., 2011; Hiller et al., 2014). These studies recruited participants from the same age group (range: 3–6 years), but differed in their choice of the children's cultural background. One exception is Hiller et al. (2014), whose focus was on a younger age group including 2-year-old children. For each study, we fitted our model to the aggregated data from all participants with varying ages.

For three ToM tasks considered in our model, the above studies revealed different orders of difficulty between cultures. Western, English-speaking children mastered the tasks in the following order: diverse beliefs, then knowledge access, and finally false beliefs. In contrast, Asian/Middle Eastern peers reversed the order between diverse beliefs and knowledge access (Note that these orders only indicate cross-sectional ToM progressions. However, Wellman et al. (2011) have recently revealed that children's longitudinal ToM progressions assessed with the ToM scale follow the same orders as obtained cross-sectionally. This validates the use of the cross-sectional data as a good approximation of the longitudinal sequences of ToM understanding for individual children). As demonstrated in **Figure 2**, our model is able to capture this cross-cultural contrast. It predicts that children's abilities to understand diverse beliefs and knowledge access multiplicatively contribute to their ability to understand false beliefs. Hence it further predicts that the order of difficulty between the former two tasks is irrelevant to the level of false belief understanding. Our fitting results confirmed these predictions, almost quantitatively reproducing both the ToM scale data from Western and Non-Western children.

Note that our model is in good agreement with the data from Hiller et al. (2014), whose focus was on a younger age group including 2-year-old children. This suggests that our model can apply to separate age groups to predict each age-related level of ToM development. We examined this possibility using ToM scale data for Japanese children (Toyama, 2007) from the only study among those listed above that reported children's ToM task performance at every age group between 3 and 6 years old. **Figure 3** shows a separate fit of our model to the data for each age group. The model fits were good, and the linear correlation between the estimated parameters (i.e., π_B, π_V, and π_{FB}) and the data over all age groups was 0.98. These results indicate that our model can capture the pattern of children's ToM abilities at each stage of their development.

Finally, we assessed whether our model can apply to children with developmental delays. Peterson et al. (2005) compared sequences of ToM development between typically developing and ToM-delayed Australian children with deafness or autism and found the same order of difficulty across all children groups, at least for three ToM tasks considered in our model. **Figure 4** shows separate fits of our model to the data for four children groups: native signers (deaf children born to signing deaf parents, mean age 10.67 years), late signers (deaf children born to non-signing hearing parents, mean age 10.01 years), autistic children (mean age 9.32 years), and typical preschoolers (mean age 4.50 years). The model fits were good even for ToM-delayed children (i.e., late singers and children with autism), suggesting that their ToM abilities, albeit delayed in development, should work in the same way as native signers and typical children.

4. DISCUSSION

We have formalized a Bayesian model of false belief reasoning that incorporates the internal models of self and others for belief formation. This model can be viewed as a version of theory-theory, explicitly representing a set of mental concepts and their interactions by a probabilistic causal model (Gopnik and Wellman, 2012). Critically, however, our model differs from the standard theory-theory in that it possesses a theory of one's own mind as well as that of other people's minds. Moreover, it allows simulation-based and standard theory-based strategies for reasoning about the belief states of others to be integrated. We have demonstrated that this hybrid approach can capture various aspects of ToM scale findings: cultural differences, age-wise development, and developmental delays with autism and deafness.

Our model predicts children's false belief ability as a multiplicative effect of their abilities to understand diverse beliefs and knowledge access. As shown above, in terms of success probabilities for corresponding ToM scale tasks, this prediction can be concisely expressed as: $\pi_{FB} \approx \pi_B \pi_V$. It is important to remember that the latter two probabilities are introduced into our model to represent the degree to which children can discern their mental states, beliefs, and visual access, from those of others. In effect, the larger these probabilities, the larger π_{FB}, and the stronger the tendency for children to recognize that others are not "like-me" in their mental states. Thus, our model predicts that developed false belief reasoning (i.e., larger π_{FB}) should rest predominantly on the internal model of others to employ a theory-based strategy and that conversely, undeveloped false belief reasoning (i.e., smaller π_{FB}) should be based mainly on the internal model of self to employ a simulation-based strategy. This differential weighting between the internal models of self and others enables our model to account for a wide variety of ToM scale data, capturing the variability of false belief ability observed across those behavioral studies.

Thus, our model is able to characterize children's competence in false belief reasoning at the various stages and aspects of their development. However, the model is not itself,

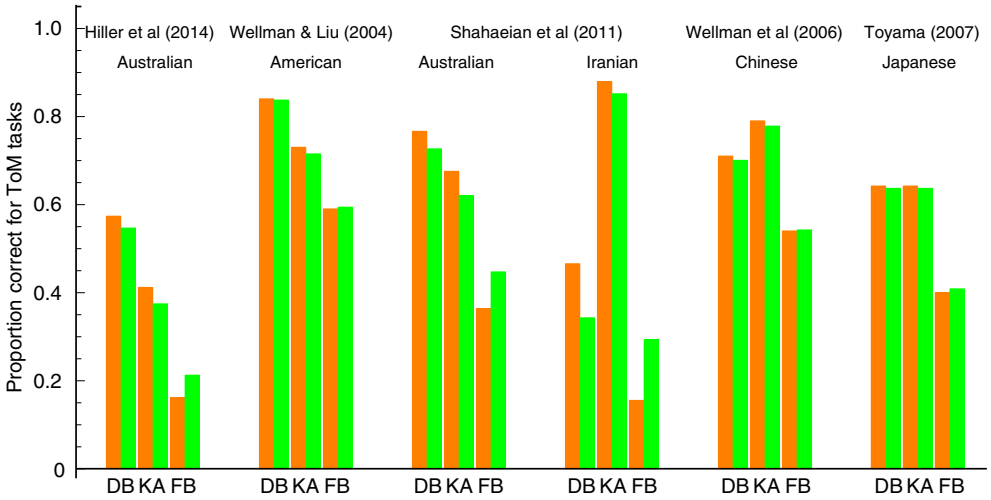


FIGURE 2 | Comparisons between our model prediction and behavioral ToM scale data for three tasks: diverse beliefs (DB), knowledge access (KA), and false beliefs (FB). Orange bars represent the behavioral data. Green bars represent the fits of our prediction (the estimated values of the parameters π_B, π_V, and π_{FB}). The left part of the figure depicts the results for Western children (Australian and American); the right part for Asian/Middle Eastern children (Iranian, Chinese, and Japanese).

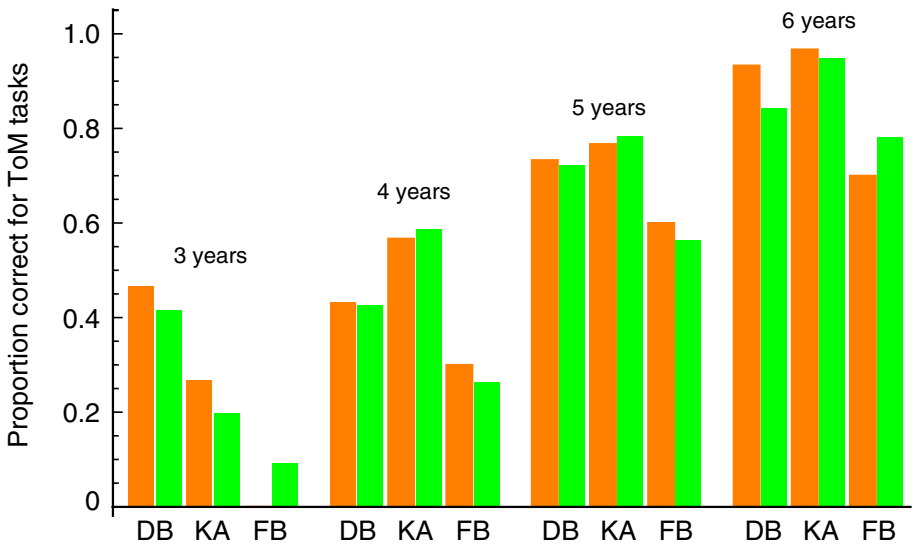


FIGURE 3 | Comparison between our model prediction and the ToM scale data of different age groups in Toyama (2007). DB, diverse beliefs; KA, knowledge access; FB, false beliefs. Orange bars represent the behavioral data. Green bars represent the fits of our prediction (the estimated values of the parameters π_B, π_V, and π_{FB}).

in its current formulation, a model for ToM acquisition. Nevertheless, it provides preliminary evidence regarding how ToM development proceeds in childhood. The key point is again the multiplicative relation: $\pi_{FB} \approx \pi_B \pi_V$. The relation states that a larger π_{FB} requires both π_B and π_V to be much larger simultaneously. Therefore, it implies that children's earlier understanding of diverse beliefs and knowledge access is a prerequisite for promoting their later false belief understanding. This naturally corresponds to a constructivist account of ToM development. Specifically, our model follows an approach of rational constructivism in cognitive development (Xu and Kushnir, 2013), as it builds on a Bayesian framework to make rational inferences. Hence, our model gives a formal constructivist interpretation of the sequential progression of ToM understandings assessed with a ToM scale.

Regarding the process of ToM development, our model makes another constructivist prediction with the multiplicative relation: $\pi_{FB} \approx \pi_B \pi_V$. A key observation is that the relation

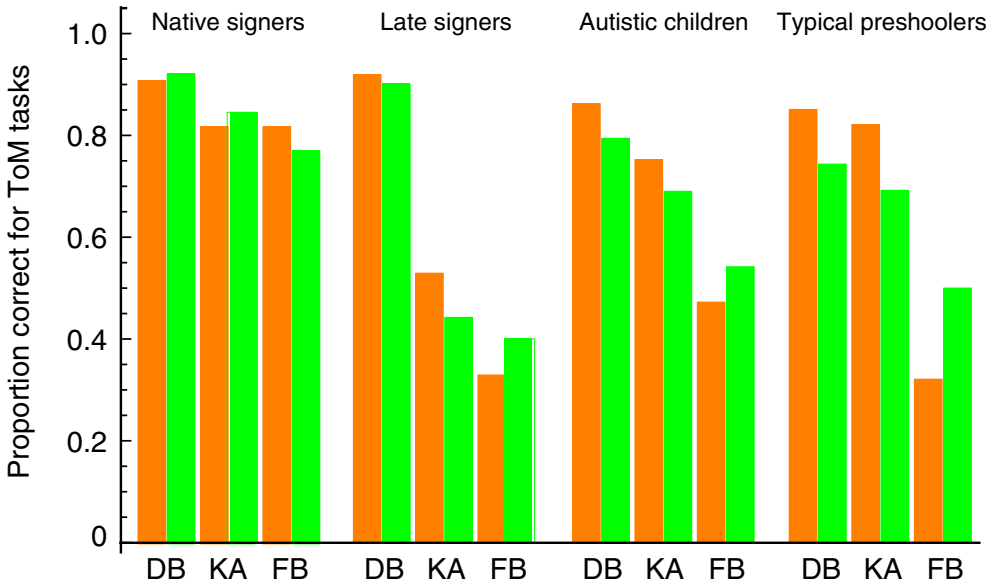


FIGURE 4 | Comparison between our model prediction and the ToM scale data of children with and without developmental delays in Peterson et al. (2005). DB, diverse beliefs; KA, knowledge access; FB, false beliefs. Orange bars represent the behavioral data. Green bars represent the fits of our prediction (the estimated values of the parameters π_B, π_V, and π_{FB}).

is bilinear: π_{FB} is linear in π_B when π_V is fixed and vice versa. This means that a fixed level of π_B or π_V affects the slope in the linear function of the other. Hence, provided that either π_B or π_V is fixed to a certain level and the other increases monotonically, a higher fixed level will result in a faster increase in the level of π_{FB}. This leads us to predict that children's initial level of understanding of diverse beliefs or knowledge access determines how fast their later understanding of false beliefs progresses over the course of ToM development. This prediction is qualitatively consistent with a recent microgenetic study by Rhodes and Wellman (2013). They demonstrated that children who had a well-developed understanding of knowledge access reliably developed an understanding of false beliefs following repeated observations of other people acting on false beliefs, whereas children who had an undeveloped understanding of knowledge access did not. Our model further predicts that children's level of diverse belief understanding also constrains their development of false belief understanding. This is due to the fact that the roles of π_B and π_V are interchangeable in our model. Pursuing this idea in a future empirical study would be worthwhile to test the prediction with microgenetic methods.

We should finally note that our Bayesian model of false beliefs, however successful, is only applicable to the unexpected-contents task, and not to the change-of-location task. To formalize false belief reasoning in the latter task, we need a related but different theory, or causal structure, to represent relevant mental state concepts. Specifically, the change-of-location task involves an extra representation of other people's actions (i.e., where to look for an object). In addition, the state of their actions depends not only on that of their beliefs (where the object is located), but also on the state of their desires (whether they want the object).

Representing this causal structure as Bayesian networks, Goodman et al. (2006) proposed two models of false beliefs: a copy theorist (CT) model and a perspective theorist (PT) model. The CT model assumes that others' beliefs depend only on the real state of the world. In contrast, the PT model assumes that they further depend on others' visual access to the world. Then it follows that the CT model is able to represent true beliefs, but is too simple to represent false beliefs, whereas the PT model is complex enough to represent both true and false beliefs. Goodman et al. modeled the development of the false belief ability as a rational transition from the CT model to the PT model: one of these models is selected for false belief reasoning according to their corresponding posterior model probabilities.

Thus, for the change-of-location task, a similar computational explanation has been advanced to understand false belief reasoning. However, similar to most behavioral ToM studies, Goodman et al. (2006) have focused on false beliefs *per se*, without taking into account extended developmental progressions of ToM leading up to false belief understanding. Therefore, we argue that, by extending our Bayesian model to accommodate the change-of-location task, it will have more explanatory power in understanding children's developing ToM abilities. Such extension is rather straightforward. It simply uses Goodman et al.'s PT model (Bayesian network) as a building block for the internal models of self and others' minds. Causal connections are then added between variables of the two internal models. The relevant variables include the state of desire as well as those of belief and visual access. This extended model, in principle, allows for Bayesian inference of other people's

action goal on the change-of-location task. Importantly, it can also cope with the unexpected-contents task since it reduces to our current formulation when irrelevant variables, in this case desire and action, are marginalized out. Furthermore, within the extended model, a causal influence between the desire states of self and others can be assessed with another task included in the ToM scale: a diverse-desires task. We are currently formalizing and validating the extended model, trying to fit it to ToM scale data including the false belief ability for the change-of-location task (Wellman and Liu, 2004; Shahaeian et al., 2014). Extending our model would thus make better use of ToM scale data to contribute to a more in-depth understanding of developmental progressions of ToM abilities.

AUTHOR CONTRIBUTIONS

NA and TI designed the study and developed the model. NA performed the model fitting to behavioral data, and prepared the manuscript. TI edited the manuscript. NA and TI discussed the results and implications of this work.

ACKNOWLEDGMENTS

This research was supported by a grant from Genesis Research Institute and a Grant-in-Aid for Scientific Research on Innovative Areas "Constructive Developmental Science" (25119503) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

Apperly, I. A. (2008). Beyond simulation-theory and theory-theory: why social cognitive neuroscience should use its own concepts to study "theory of mind". *Cognition* 107, 266–283. doi: 10.1016/j.cognition.2007.07.019

Berthiaume, V. G., Shultz, T. R., and Onishi, K. H. (2013). A constructivist connectionist model of transitions on false-belief tasks. *Cognition* 126, 441–458. doi: 10.1016/j.cognition.2012.11.005

Gallese, V., and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends Cogn. Sci.* 2, 493–501. doi: 10.1016/S1364-6613(98)01262-5

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York, NY: Oxford University Press. doi: 10.1093/0195138929.001.0001

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., et al. (2006). "Intuitive theories of mind: a rational approach to false belief," in *Proceedings of the 28th Annual Conference of Cognitive Science Society*, eds R. Sun and N. Miyake (Mahwah, NJ: Lawrence Erlbaum Associates), 1382–1387.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., and Kushnir, T. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* 111, 3–32. doi: 10.1037/0033-295X.111.1.3

Gopnik, A., and Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Dev. Sci.* 10, 281–287. doi: 10.1111/j.1467-7687.2007.00584.x

Gopnik, A., and Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind Lang.* 7, 145–171. doi: 10.1111/j.1468-0017.1992.tb00202.x

Gopnik, A., and Wellman, H. M. (1994). "The theory theory," in *Mapping the Mind*, eds L. Hirschfeld and S. Gelman (New York, NY: Cambridge University Press), 257–293.

Gopnik, A., and Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms and the theory theory. *Psychol. Bull.* 138, 1085–1108. doi: 10.1037/a0028044

Gordon, R. M. (1986). Folk psychology as simulation. *Mind Lang.* 1, 158–171. doi: 10.1111/j.1468-0017.1986.tb00324.x

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). "Bayesian models of cognition," in *The Cambridge Handbook of Computational Psychology*, ed R. Sun (Cambridge: Cambridge University Press), 59–100. doi: 10.1017/CBO9780511816772.006

Griffiths, T. L., and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cogn. Psychol.* 51, 334–384. doi: 10.1016/j.cogpsych.2005.05.004

Hiller, R. M., Weber, N., and Young, R. L. (2014). The validity and scalability of the theory of mind scale with toddlers and preshoolers. *Psychol. Assess.* 26, 1388–1393. doi: 10.1037/a0038320

Jacobs, R. A., and Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdiscip. Rev.* 2, 8–21. doi: 10.1002/wcs.80

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as bayesian inference. *Annu. Rev. Psychol.* 55, 271–304. doi: 10.1146/annurev.psych.55.090902.142005

Leslie, A. M., Friedman, O., and German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends Cogn. Sci.* 8, 528–533. doi: 10.1016/j.tics.2004.10.001

Lu, H., Yuille, A. L., Liljeholm, M., Chang, P. W., and Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychol. Rev.* 115, 955–984. doi: 10.1037/a0013256

Mahy, C. E., Moses, L. J., and Pfeifer, J. H. (2014). How and where: theory-of-mind in the brain. *Dev. Cogn. Neurosci.* 9, 68–81. doi: 10.1016/j.dcn.2014.01.002

Marr, D. (1982). *Vision*. New York, NY: W.H. Freeman.

Meltzoff, A. N. (2007a). 'Like me': a foundation for social cognition. *Dev. Sci.* 10, 126–134. doi: 10.1111/j.1467-7687.2007.00574.x

Meltzoff, A. N. (2007b). The 'like me' framework for recognizing and becoming an intentional agent. *Acta Psychol.* 124, 26–43. doi: 10.1016/j.actpsy.2006.09.005

Mitchell, P., Currie, G., and Ziegler, F. (2009). Two routes to perspective: simulation and rule-use as approaches to mentalizing. *Brit. J. Dev. Psychol.* 27, 513–514. doi: 10.1348/026151008X334737

Nichols, S., and Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

O'Laughlin, C., and Thagard, P. (2000). Autism and coherence: a computational model. *Mind Lang.* 15, 375–392. doi: 10.1111/1468-0017.00140

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York, NY: Cambridge University Press.

Perner, J., Leekam, S. R., and Wimmer, H. (1987). Three-year-olds' difficulty with false belief: the case for a conceptual deficit. *Brit. J. Dev. Psychol.* 5, 125–137. doi: 10.1111/j.2044-835X.1987.tb01048.x

Peterson, C. C., Wellman, H. M., and Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Dev.* 76, 502–517. doi: 10.1111/j.1467-8624.2005.00859.x

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526. doi: 10.1017/S0140525X00076512

Rhodes, M., and Wellman, H. (2013). Constructing a new theory from old ideas and new evidence. *Cogn. Sci.* 37, 592–604. doi: 10.1111/cogs.12031

Saxe, R. (2005). Against simulation: the argument from error. *Trends Cogn. Sci.* 9, 174–179. doi: 10.1016/j.tics.2005.01.012

Shahaeian, A., Nielsen, M., Peterson, C. C., and Slaughter, V. (2014). Cultural and family influences on children's theory of mind development: a comparison of australian and iranian school-age children. *J. Cross Cult. Psychol.* 45, 555–568. doi: 10.1177/0022022113513921

Shahaeian, A., Peterson, C. C., Slaughter, V., and Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Dev. Psychol.* 47, 1239–1247. doi: 10.1037/a0023899

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search, 2nd Edn*. Cambridge: MIT Press.

Toyama, K. (2007). Examining theory-of-mind tasks with japanese children: the Wellman and Liu tasks. *Jpn. J. Educ. Psychol.* 55, 359–369. doi: 10.5926/jjep1953.55.3_359

Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev.* 72, 655–684. doi: 10.1111/1467-8624.00304

Wellman, H. M., Fang, F., Liu, D., Zhu, L., and Liu, G. (2006). Scaling of theory-of-mind understandings in chinese children. *Psychol. Sci.* 17, 1075–1081. doi: 10.1111/j.1467-9280.2006.01830.x

Wellman, H. M., Fang, F., and Peterson, C. C. (2011). Sequential progressions in a theory of mind scale: longitudinal perspectives. *Child Dev.* 82, 780–792. doi: 10.1111/j.1467-8624.2011.01583.x

Wellman, H. M., and Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Dev.* 75, 523–541. doi: 10.1111/j.1467-8624.2004.00691.x

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13, 103–128. doi: 10.1016/0010-0277(83)90004-5

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 593–602. doi: 10.1098/rstb.2002.1238

Xu, F., and Kushnir, T. (2013). Infants are rational constructivist learners. *Curr. Direct. Psychol. Sci.* 22, 28–32. doi: 10.1177/0963721412469396

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Fitting the Model of False Belief Reasoning to ToM Scale Data

We adopted a Bayesian method for estimating the parameters π_B, π_V, δ, and ϵ to fit Equation (14) to ToM scale data D for three tasks: diverse beliefs, knowledge access, and unexpected-contents false beliefs. This can be done through maximizing the posterior probability distribution for the parameters:

$$P(\pi_B, \pi_V, \delta, \epsilon|D) = \frac{P(D|\pi_B, \pi_V, \delta, \epsilon)P(\pi_B, \pi_V, \delta, \epsilon)}{P(D)} \quad \text{(A1)}$$

where $P(D|\pi_B, \pi_V, \delta, \epsilon)$ is the likelihood function of the data and $P(\pi_B, \pi_V, \delta, \epsilon)$ is the prior probability distribution for the parameters. $P(D)$ is a normalization constant. This is due to Bayes' rule. We assume an individual child's responses to the three tasks as independent Bernoulli trials with success probabilities π_B for diverse beliefs, π_V for knowledge access, and π_{FB} for unexpected-contents false beliefs. We can then represent the number of children passing each task using the sum of the Bernoulli trials with the values of 1 for success and 0 for failure. This sum has a binomial distribution, giving the likelihood function of the data for each task. The joint likelihood function of the entire data set is the product of three likelihood functions for each task and is given by:

$$\begin{aligned} L(\pi_B, \pi_V, \pi_{FB}; D) = {} & {}_n\mathrm{C}_{r_B}{\pi_B}^{r_B}(1-\pi_B)^{n-r_B} \cdot {}_n\mathrm{C}_{r_V}{\pi_V}^{r_V} \\ & (1-\pi_V)^{n-r_V} \cdot {}_n\mathrm{C}_{r_{FB}}{\pi_{FB}}^{r_{FB}}(1-\pi_{FB})^{n-r_{FB}} \end{aligned} \quad \text{(A2)}$$

where n is the number of participating children and r_B, r_V and r_{FB} are the numbers of children passing the corresponding tasks. The likelihood function $P(D|\pi_B, \pi_V, \delta, \epsilon)$ is then obtained by substituting Equation (14) into π_{FB}. For the prior probability distribution $P(\pi_B, \pi_V, \delta, \epsilon)$, we assume uniform priors on π_B and π_V, and asymmetric beta priors on δ and ϵ with a beta distribution: Beta(2, 48) (mode 0.02; mean 0.04; variance 0.00075). Thus, the prior probability reduces to $P(\delta)P(\epsilon) = \mathrm{Beta}(2, 48) \cdot \mathrm{Beta}(2, 48)$.

As shown above, the posterior distribution is proportional to the product of the likelihood and the prior distribution. For convenience, we numerically maximized the logarithm of the product with respect to the parameters to obtain their maximum a posteriori estimates. We did this using Mathematica's built-in NMaximize function with the constraint that each parameter takes values between zero and one (i.e., all parameters should be probabilities). Substituting these estimates into Equation (14), we obtained a prediction of π_{FB}.